Traitement automatique des langues peu dotées

Delphine Bernhard[*] — **Claudia Soria**[**]

[*] *LiLPa, Université de Strasbourg*
[**] *CNR-ILC, Pisa*

RÉSUMÉ. *Jusqu'à récemment, la plupart des travaux de recherche sur le traitement automatique des langues (TAL) étaient axés sur quelques langues bien décrites avec de nombreux locuteurs. La situation évolue rapidement, avec une nette augmentation de l'intérêt pour les langues dites « sous-dotées ». L'objectif de ce numéro de la revue* Traitement automatique des langues *est de donner un aperçu des recherches en cours sur le TAL pour des langues peu dotées du monde entier, couvrant une grande variété de tâches. Les articles sélectionnés portent à la fois sur des langues qui sont encore au début du processus et sur des langues dont la situation s'est très récemment améliorée. Nous espérons qu'ils pourront servir à orienter les recherches futures sur d'autres langues disposant de peu de ressources et d'outils.*

ABSTRACT. *Until recently, most of the research work in Natural Language Processing (NLP) has been focused on a few well-described languages with many speakers. The situation is rapidly evolving, with a clear increase in the interest towards so called "under-resourced" languages. The goal of this issue of the Traitement Automatique des Langues journal is to give an overview of current research on NLP for under-resourced languages from all over the world, encompassing a large variety of tasks. The selected papers address languages which are still at very early stages as well as languages whose situation has very recently improved. We hope that they can be helpful to guide future research on other languages with little or no resources and tools.*

MOTS-CLÉS : *langues peu dotées, ressources linguistiques, outils de TAL.*

KEYWORDS: *under-resourced languages, language resources, NLP tools.*

1. Introduction

Jusqu'à récemment, la plupart des travaux de recherche en traitement automatique des langues (TAL) se sont concentrés sur quelques langues bien décrites et ayant de nombreux locuteurs (Del Gratta *et al.*, 2014). Le manque d'intérêt pour d'autres langues et variétés linguistiques «sous-dotées» peut s'expliquer par différentes raisons : manque de financement, de ressources humaines, de technologie appropriée, de descriptions linguistiques complètes et précises, de reconnaissance académique par la communauté scientifique, pour n'en nommer que quelques-unes. Les langues sous-dotées posent néanmoins d'importants défis scientifiques qui ouvrent des pistes de progrès pour le TAL en général.

Premièrement, à une époque où les méthodes de l'état de l'art nécessitent généralement de grandes quantités de données annotées, le travail sur des langues sous-dotées impose souvent des méthodes capables de traiter des jeux de données de petite taille (*small data*). Par exemple, les plus petits corpus arborés issus des Universal Dependencies ne contiennent que quelques milliers de *tokens* : environ 1 000 pour l'akkadien ou le sanskrit, 10 000 pour le breton ou le féroïen (pour ne citer que quelques langues), contre plus de 1 million pour l'arabe ou le français (Universal Dependencies, 2018). Des méthodes efficaces et fiables pour l'acquisition et la collecte de ressources et d'annotations (OCR, *crowdsourcing*, médias sociaux, etc.) sont essentielles (McShane *et al.*, 2002). De plus, les outils d'annotation automatique doivent être capables de gérer le manque de données (Abraham *et al.*, 2016), les problèmes de qualité et les mots hors vocabulaire (Liu et Kirchhoff, 2018). Ces dernières années, des modèles s'appuyant sur la projection d'annotations à partir d'autres langues ont vu le jour (Sukhareva et Chiarcos, 2014 ; Agić *et al.*, 2016). Ces méthodes tirent parti des outils d'annotation existants ou des ressources annotées pour des langues mieux dotées, grâce à l'utilisation de corpus parallèles. Elles dépendent donc largement de la qualité de l'alignement que l'on peut obtenir (Akbik et Vollgraf, 2018).

Deuxièmement, compte tenu des difficultés à trouver des ressources telles que des lexiques ou des corpus, les données collectées sont souvent très hétérogènes et correspondent à différentes époques, aires linguistiques ou domaines, par exemple des corpus de textes intégrant différentes variétés géolinguistiques et portant sur différents sujets à différentes époques (Goldhahn *et al.*, 2016). Cette hétérogénéité implique aussi souvent des variations dans la graphie, dues soit à une évolution des normes orthographiques dans le temps, soit à l'absence de normes orthographiques pour les langues ou les variétés linguistiques qui sont essentiellement orales et rarement écrites (Kurimo *et al.*, 2017). Dans de tels cas, la normalisation orthographique est souvent la solution préférée (Samardzic *et al.*, 2015). En outre, l'alternance codique peut également causer des problèmes, ce qui nécessite d'identifier la langue ou la variété de langue (Lavergne *et al.*, 2014).

Troisièmement, les travaux de TAL pour les langues sous-dotées ont tendance à être réalisés dans des groupes de recherche isolés ou dispersés, et les ressources produites utilisent souvent des formats et des normes différents. Trouver ces ressources, y

accéder et les rendre interopérables pour qu'elles puissent être réutilisées peut devenir un défi en soi. Quand il s'agit de langues sous-dotées, les questions d'interopérabilité des données et des métadonnées deviennent d'une importance cruciale pour combiner et réutiliser les quelques ressources et outils qui pourraient être disponibles (Alegria *et al.*, 2011).

L'objectif de ce numéro de *Traitement automatique des langues* est de donner un aperçu de la recherche actuelle sur le TAL pour les langues peu dotées du monde entier, englobant une grande variété de tâches. Nous avons reçu 23 soumissions en tout, ce qui montre que le sujet abordé dans ce numéro spécial est particulièrement pertinent et figure en bonne place parmi les thématiques de recherche actuelles. Il ne s'agit plus d'un sujet de recherche très spécialisé et secondaire. C'est ce qu'attestent également les nombreuses conférences et ateliers consacrés au TAL pour les langues peu dotées ces dernières années : par exemple la série d'ateliers CCURL (*Collaboration and Computing for Under-Resourced Languages*) à partir de 2014, SLTU (*Spoken Language Technologies for Under-Resourced Languages*) à partir de 2008 ou le « *Less-Resourced Languages Workshop* » organisé depuis 2009 à la conférence L&TC. Les principales conférences du domaine, telles que COLING ou ACL, proposent régulièrement des sessions dédiées aux langues moins dotées. En 2017, un groupe d'intérêt spécial ELRA-ISCA sur les langues sous-dotées (SIGUL) a été créé. Les défis posés par les langues peu dotées sont de plus en plus souvent abordés dans les communications présentées lors de conférences et d'ateliers généralistes, en particulier pour évaluer les méthodes face au manque de données, par exemple « *CoNLL 2018 UD Shared Task* » (Zeman *et al.*, 2018), la campagne d'évaluation IWSLT 2018 (Jan *et al.*, 2018), ou l'atelier « *Workshop on Deep Learning Approaches for Low-Resource NLP* » à ACL 2018 (Haffari *et al.*, 2018). Les langues peu dotées fournissent en effet des données d'évaluation difficiles, mais réalistes, pour les méthodes état de l'art.

Plus important encore, le fait que des travaux de recherche universitaire de grande qualité soient menés pour les langues peu dotées pourrait être lié à une prise de conscience partagée et de plus en plus répandue de l'importance de la représentation numérique pour toutes les langues. Alors que la numérisation de la société moderne s'accentue et que la fracture numérique entre les langues très utilisées et les langues moins répandues se creuse, les locuteurs de langues minoritaires ou moins répandues sont confrontés à des disparités dans l'accès à l'information. La disponibilité de contenus dans de nombreuses langues différentes réduirait certainement cette inégalité.

Le développement des technologies TAL et des ressources linguistiques pour les langues qui ont été ou sont encore exclues du monde numérique est une condition préalable pour qu'elles soient pleinement utilisables sur les médias numériques dans un avenir que l'on n'espère pas si lointain. Cela garantira une meilleure égalité des droits numériques à tous les citoyens, qui pourraient ainsi bénéficier d'un environnement favorable pour pouvoir s'exprimer et créer leur propre contenu culturel dans les langues locales, un pas de plus vers une diversité linguistique et culturelle plus large et plus forte.

2. Résumés des articles

Le présent numéro se compose de quatre articles traitant de langues à différents stades de développement concernant les ressources et les outils pour le TAL.

Les deux premiers articles se concentrent sur l'acquisition de données annotées, en s'appuyant sur deux approches différentes : une campagne d'annotation « classique » réalisée par des annotateurs formés pour l'annotation du serbe, et une approche reposant sur le *crowdsourcing* pour deux langues de France, l'alsacien et le créole guadeloupéen.

Le troisième article porte sur l'analyse syntaxique de deux langues à faibles ressources, le same du nord et le komi-zyriène.

Enfin, le dernier article donne un aperçu de l'état de l'art du traitement automatique du dialecte tunisien, qui recense les ressources et outils disponibles.

2.1. *De la constitution d'un corpus arboré à l'analyse syntaxique du serbe*

L'article décrit le processus de production d'un corpus d'environ 101 000 *tokens* en serbe, annoté avec des propriétés morphosyntaxiques, les lemmes et les relations de dépendance syntaxique. Avant les travaux décrits dans l'article, le serbe a été doté de certaines ressources et outils, mais la plupart d'entre eux ne sont pas librement et facilement disponibles (à l'exception notable de certains travaux récents : un lexique morphosyntaxique, publié en 2016, et le corpus Universal Dependencies qui a été publié en 2017). La méthodologie proposée dans cet article vise à optimiser les ressources linguistiques et humaines limitées en adaptant les ressources existantes d'une langue très proche (le croate), en utilisant des ressources lexicales qui ont été produites de manière collaborative (Wiktionary) et en préannotant automatiquement le corpus.

2.2. *À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées*

L'utilisation du *crowdsourcing* pour l'annotation est encore relativement inexplorée pour les langues sous-dotées. Dans cet article, une méthodologie d'annotation en parties du discours s'appuyant sur le *crowdsourcing* est présentée et appliquée aux dialectes alsaciens et au créole guadeloupéen. Dans les deux cas, les ressources existantes sont très limitées et la graphie n'est pas normalisée. Deux plateformes participatives différentes sont présentées. La première vise à obtenir des annotations en parties du discours. Les corpus sont préannotés avant d'être présentés aux participants. Les corpus qui en résultent sont ensuite utilisés pour entraîner des étiqueteurs. La deuxième plateforme vise la collecte de corpus bruts (en se concentrant pour l'instant sur les recettes de cuisine) ainsi que des lexiques de variantes graphiques et dialectales. Les utilisateurs peuvent également corriger des annotations morphosyntaxiques qui ont été produites automatiquement.

2.3. *Analyse syntaxique de langues faiblement dotées à partir de plongements de mots multilingues*

Le same du nord et le komi-zyriène sont deux langues finno-ougriennes sous-dotées qui n'ont pas les mêmes niveaux de développement en ce qui concerne les ressources et les outils de TAL : alors que le same du nord dispose de lexiques flexionnels relativement complets, ainsi que d'un corpus Universal Dependencies, le komi-zyriène manque de telles ressources. L'approche proposée pour l'analyse automatique de ces langues est multilingue et ne nécessite qu'un petit lexique bilingue et une annotation syntaxique manuelle de quelques phrases. La méthode utilise également des plongements de mots, à la fois pour les langues cibles et pour des langues mieux dotées en ressources (finnois ou russe), ainsi que des corpus Universal Dependencies existants (anglais, finnois ou russe).

2.4. *Un état de l'art du traitement automatique du dialecte tunisien*

Le dialecte tunisien est l'un des nombreux dialectes parlés dans les pays arabes. Il diffère considérablement de l'arabe standard moderne, qui a été beaucoup plus étudié et doté de nombreuses ressources et outils. L'article passe en revue les travaux récents visant à collecter des ressources et des outils pour le dialecte tunisien. Il se concentre sur les aspects suivants : corpus (transcriptions orales, Web, corpus parallèles), lexiques, ontologies, traitement de la parole, outils d'analyse morpho-syntaxique, identification de la langue, traduction, analyse des sentiments, normalisation.

La plupart des travaux décrits sont très récents, ce qui montre que la communauté des chercheurs s'est récemment beaucoup intéressée au développement de ressources et d'outils pour le dialecte tunisien. Cette implication a permis une nette amélioration de la situation. Cependant, l'article note que seulement 24 % des ressources énumérées sont téléchargeables gratuitement en ligne, et uniquement deux outils. De plus, les ressources sont encore assez petites et souvent limitées à un domaine spécifique. En conséquence, l'effort de construction de ressources et d'outils pour le dialecte tunisien doit se poursuivre.

3. Conclusion

Ce numéro est consacré aux langues peu et sous-dotées. Ces termes n'ont pas encore de définition précise et se recoupent largement avec ceux des langues minoritaires et en danger. Ce qui est clair, c'est que ces termes peuvent s'appliquer à un large éventail de langues, à différents stades d'avancement en ce qui concerne les ressources et les outils du TAL. Les langues sous-dotées peuvent être minoritaires ou menacées, mais l'inverse ne s'applique pas toujours (par exemple, le catalan est une langue minoritaire en Espagne, mais ne dispose pas de moins de ressources). D'un autre côté, les langues comptant des millions de locuteurs, et vitales comme l'ourdou

ou certaines langues chinoises, sont également sous-dotées. Appliquer le terme à une langue implique de savoir si la langue considérée dispose ou non des ressources et de la technologie nécessaires pour accéder aux médias numériques comme les autres.

Deux articles de ce numéro traitent de langues qui n'en sont qu'à leurs débuts (l'alsacien, le créole guadeloupéen, le same du nord, le komi-zyriène) et peuvent être utiles pour orienter les recherches futures sur d'autres langues avec peu ou pas de ressources. Les deux autres articles concernent des langues qui ne peuvent probablement plus être considérées comme des langues peu dotées (serbe, dialecte tunisien), mais qui ont très récemment amélioré leur situation. Tous ces exemples montrent comment le statut d'une langue peut non seulement être amélioré, mais aussi conduire à des solutions innovantes. Nous espérons que ces articles intéresseront non seulement les chercheurs, mais aussi les institutions qui prennent des décisions en matière de financement et de politiques linguistiques, afin d'encourager la recherche sur les nombreuses langues qui manquent encore de ressources.

Remerciements

Nous aimerions remercier les rédacteurs en chef et le comité de rédaction de la revue TAL d'avoir eu l'idée d'un numéro spécial sur le thème des langues peu dotées. Nous remercions également les relectrices et relecteurs pour leurs précieux commentaires.

Membres du comité scientifique (par ordre alphabétique) : Gilles Adda (LIMSI-CNRS, France), Antti Arppe (University of Alberta, Canada), Vincent Berment (INALCO, France), Myriam Bras (Université Toulouse Jean Jaurès / CLLE-ERSS, France), Thierry Declerck (DFKI, Allemagne), Chantal Enguehard (LS2N, Nantes, France), Vera Ferreira (CIDLeS - Interdisciplinary Centre for Social and Language Documentation, Portugal), Karën Fort (Université Paris-Sorbonne, France), András Kornai (Hungarian Academy of Sciences, Hongrie), Anne-Laure Ligozat (ENSIIE / LIMSI-CNRS, France), Teresa Lynn (ADAPT DCU, Irlande), Mathieu Mangeot-Nagata (Université de Savoie / LIG, France), Joseph Mariani (LIMSI-CNRS, France), Damien Nouvel (INALCO), Thierry Poibeau (LATTICE-CNRS, France), Laurette Pretorius (University of South Africa, Afrique du Sud), Benoît Sagot (Inria Paris, France), Sakriani Sakti (NAIST, Japon), Kevin Scannell (Saint Louis University, Missouri, Etats-Unis), Yves Scherrer (University of Helsinki, Finlande), Jörg Tiedemann (University of Helsinki, Finlande), Trond Trosterud (Tromsø University, Norvège), Francis Tyers (Moscow Higher School of Economics, Russie), Assaf Urieli (Université Toulouse Jean Jaurès / CLLE-ERSS et Joliciel Informatique, France).

4. Bibliographie

Abraham B., Umesh S., Joy N., « Overcoming Data Sparsity in Acoustic Modeling of Low-Resource Language by Borrowing Data and Model Parameters from High-Resource Languages », *Proceedings of the 17th Annual Conference of the International Speech Communi-*

cation Association (INTERSPEECH 2016) : Understanding Speech Processing in Humans and Machines, p. 3037-3041, 2016.

Agić Ž., Johannsen A., Plank B., Alonso H. M., Schluter N., Søgaard A., « Multilingual Projection for Parsing Truly Low-Resource Languages », *Transactions of the Association for Computational Linguistics*, vol. 4, p. 301-312, 2016.

Akbik A., Vollgraf R., « ZAP : An Open-Source Multilingual Annotation Projection Framework », *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Alegria I., Artola X., de Ilarraza A. D., Sarasola K., « Strategies to develop Language Technologies for Less-Resourced Languages based on the case of Basque », *HAL Id : artxibo-00783393*, 2011.

Del Gratta R., Frontini F., Khan A. F., Mariani J., Soria C., « The LREMap for Under-Resourced Languages », *Proceedings of CCURL 2014 : Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, p. 78-83, 2014.

Goldhahn D., Sumalvico M., Quasthoff U., « Corpus collection for under-resourced languages with more than one million speakers », *Proceedings of the LREC Workshop "CCURL 2016 Collaboration and Computing for Under-Resourced Languages : Towards an Alliance for Digital Language Diversity"*, Portorož, Slovenia, 2016.

Haffari R., Cherry C., Foster G., Khadivi S., Salehi B., *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, 2018.

Jan N., Cattoni R., Sebastian S., Cettolo M., Turchi M., Federico M., « The IWSLT 2018 Evaluation Campaign », *Proceedings of the International Workshop on Spoken Language Translation*, p. 2-6, 2018.

Kurimo M., Enarvi S., Tilk O., Varjokallio M., Mansikkaniemi A., Alumaë T., « Modeling under-resourced languages for speech recognition », *Language Resources and Evaluation*, vol. 51, n° 4, p. 961-987, 2017.

Lavergne T., Adda G., Adda-Decker M., Lamel L., « Automatic Language Identity Tagging on Word and Sentence-Level in Multilingual Text Sources : a Case-Study on Luxembourgish », *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

Liu A., Kirchhoff K., « Context Models for OOV Word Translation in Low-Resource Languages », *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1 : Research Papers)*, p. 54-67, 2018.

McShane M., Nirenburg S., Cowie J., Zacharski R., « Embedding knowledge elicitation and MT systems within a single architecture », *Machine Translation*, n° 17, p. 271-305, 2002.

Samardzic T., Scherrer Y., Glaser E., « Normalising orthographic and dialectal variants for the automatic processing of Swiss German », *Proceedings of the 7th Language and Technology Conference*, Poznan, 2015.

Sukhareva M., Chiarcos C., « Diachronic proximity vs. data sparsity in cross-lingual parser projection. A case study on Germanic », *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, p. 11-20, 2014.

Universal Dependencies, `http://universaldependencies.org/`, 2018. Accédé le 9 décembre 2018.

Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S., « CoNLL 2018 shared task : Multilingual parsing from raw text to universal dependencies », *Procee-*

dings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies, p. 1-21, 2018.